AlexUNLP-STM at NADI 2024 shared task: Quantifying the Arabic Dialect Spectrum with Contrastive Learning, Weighted Sampling, and BERT-based Regression Ensemble

Abdelrahman Sakr, Marwan Torki, and Nagwa El-Makky
Computer and Systems Engineering Department, Alexandria University
{es-abdelrahmansakr2023, mtorki, nagwamakky}@alexu.edu.eg

Abstract

Recognizing the nuanced spectrum of dialectness in Arabic text poses a significant challenge for natural language processing (NLP) tasks. Traditional dialect identification (DI) methods treat the task as binary, overlooking the continuum of dialect variation present in Arabic speech and text. In this paper, we describe our submission to the NADI shared Task of ArabicNLP 2024. We participated in Subtask 2 - ALDi Estimation, which focuses on estimating the Arabic Level of Dialectness (ALDi) for Arabic text, indicating how much it deviates from Modern Standard Arabic (MSA) on a scale from 0 to 1, where 0 means MSA and 1 means high divergence from MSA. We explore diverse training approaches, including contrastive learning, applying a random weighted sampler along with fine-tuning a regression task based on the AraBERT model, after adding a linear and non-linear layer on top of its pooled output. Finally, performing a brute force ensemble strategy increases the performance of our system. Our proposed solution achieved a Root Mean Squared Error (RMSE) of 0.1406, ranking second on the leaderboard.

1 Introduction

The Arabic language exhibits a duality of forms: Modern Standard Arabic (MSA) serves as the standardized variety taught in schools and used in formal settings and communication across the Arab world. In contrast, numerous regional varieties of Dialectal Arabic (DA) dominate everyday spoken communication, including interactions on social media. These dialects significantly differ from MSA in phonology, grammar, and vocabulary (Habash, 2010).

This bifurcation of the language, coupled with frequent code-switching between MSA and DA, presents significant challenges for Arabic NLP systems (Sakr and Torki, 2023). Researchers have developed numerous systems to address Dialect Identification (DI), often at the sentence level (Zaidan and Callison-Burch, 2011; Elfardy and Diab, 2013; Salameh et al., 2018) and occasionally at the token level to identify instances of code-switching (Solorio et al., 2014; Molina et al., 2016). However, these approaches often adopt a binary perspective, categorizing a sentence or token as either MSA or DA.

To tackle these challenges, (Keleg et al., 2023) introduced the AOC-ALDi dataset, which contains 127,835 sentences labeled for their level of dialectal variation, derived from the AOC dataset (Zaidan and Callison-Burch, 2011). The AOC dataset includes user comments from three newspapers, representing Egyptian, Levantine, and Gulf dialects.

In this paper, we explore the use of pre-trained transformer-based language models. Given that the AOC-ALDi dataset is imbalanced, with approximately 50% consisting of MSA samples labeled as 0, transformer models often overfit in this context. Therefore, we fine-tuned the linear and non-linear layers on top of these pre-trained language models and explored different training strategies, including contrastive learning and the use of a weighted random sampler in the data loader. Combining these approaches resulted in further improvements in system performance. Finally, employing a brute-force ensemble strategy with the collaboration of 5 checkpoints from different training paradigms further increased the effectiveness of our system.

Throughout this paper, we use the following abbreviations: Weighted Random Sampler (WRS), Contrastive Loss (CL), Temperature Parameter for NT-Xent loss (T), Remove Non-Arabic Letters and Links (RMNALL), and Drop 43,000 Samples from Majority Label (DS).

2 Related Work

Early research on Arabic Dialect Identification (DI) focused on both binary MSA-DA classification and

Proceedings of The Second Arabic Natural Language Processing Conference, pages 735–741

multi-class problems involving various DA variants (Althobaiti, 2020; Keleg and Magdy, 2023). Arabic DI has garnered significant attention, leading to multiple shared tasks (Zampieri et al., 2014; Abdul-Mageed et al., 2020) and datasets (Zaidan and Callison-Burch, 2011; Abdelali et al., 2021; Althobaiti, 2022). Research has been conducted at both the sentence and token levels, but existing methods often struggle to differentiate between sentences that share similar dialectal cues but exhibit varying levels of dialectness. To address these issues, (Keleg et al., 2023) introduced the AOC-ALDi dataset, which includes 127,835 sentences labeled for dialectness, derived from user comments on articles from three newspapers representing Egyptian, Levantine, and Gulf dialects (Zaidan and Callison-Burch, 2011). They employed a BERT-based regression model, fine-tuning a regression head atop MarBERT (Abdul-Mageed et al., 2021a), and conducted multiple rounds of fine-tuning with different random seeds to ensure consistency.

There has been an interest in contrastive learning (Chen et al., 2020) in recent years due to its significant advancements in various NLP tasks. However, its application within the Arabic context has been relatively limited. This trend has shifted as demonstrated by (Shapiro et al., 2022), who showcased the efficacy of contrastive learning in Arabic hate speech detection, resulting in substantial improvements over baseline methods.

Recent studies in computer vision (Pintea et al., 2023) have shown that incorporating a classification loss alongside a regression loss enhances the performance of deep regression models. They found that this approach is particularly beneficial for regression tasks with imbalanced data distributions.

3 Data

For our experiments, we used the training AOC-ALDi dataset provided by the shared task. We conducted an extensive analysis of the dataset to gain insights that could aid us in effectively addressing the problem. The dataset comprises 102,886 data samples with 68 unique labels ranging from 0 to 1 where 0 represents MSA, and 1 signifies high divergence from MSA. The distribution of the dataset is illustrated in Figure 1.

We observed a significant imbalance in the dataset, where the MSA label predominates over other labels. Specifically, there are 53,844 data samples labeled as 0 (MSA), accounting for approximately 52.3% of the total dataset. To mitigate this imbalance, we employed WRS in the data loader. Additionally, we explored undersampling the majority class by removing around 43,000 samples from the MSA data samples.

We also performed simple data preprocessing by removing non-Arabic letters from the sentences, but it decreased the system's performance. Finally, we split this dataset into an 85% training set and a 15% dev set. Our dev set serves to monitor the training process and detect issues such as overfitting or underfitting and to keep the best checkpoints of the training procedure.

For the shared task provided dev set, we found that it has 107 data samples, which have a slightly different distribution from the provided training dataset. The distribution of the shared task dev set can be found in Figure 2 for this circumstance, we saved the top 4 checkpoints of any training paradigm on our dev set and then tested them all on the subtask 2 dev set to choose the best one.

4 System

4.1 Encoders Selection

We use the following BERT-based models in our experiments:

4.1.1 AraBERTv2-Twitter

AraBERT (Antoun et al.) is based on Google's BERT (Devlin et al., 2019) architecture. It is an Arabic pre-trained language model. There are multiple versions of this model. We chose AraBERTv2-Twitter[1] because it is pre-trained on 77 GB of MSA text plus 60 million Arabic dialect tweets. We use the Large[2] version of it, which has more parameters, totaling 371 million.

4.1.2 MARBERTv2

MARBERTv2 (Abdul-Mageed et al., 2021a) is a large-scale Arabic language model that focuses also on MSA and different dialects. We chose MARBERTv2[3] because it is trained on 1 billion Arabic tweets, similar to MARBERTv1[4] (Abdul-Mageed et al., 2021b), in addition to the same MSA texts as ARBERT in addition to AraNews dataset with a bigger sequence length of 512 tokens.

[1] https://huggingface.co/aubmindlab/bert-base-arabertv02-twitter

[2] https://huggingface.co/aubmindlab/bert-large-arabertv02

[3] https://huggingface.co/UBC-NLP/MARBERTv2

[4] UBC-NLP/MARBERT

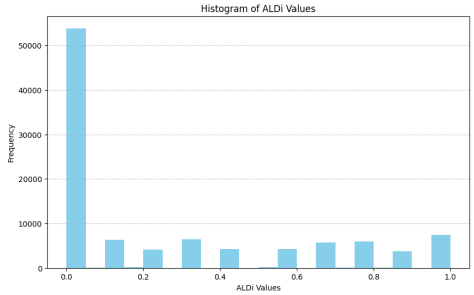

Figure 1: Provided training AOC-ALDi dataset distribution

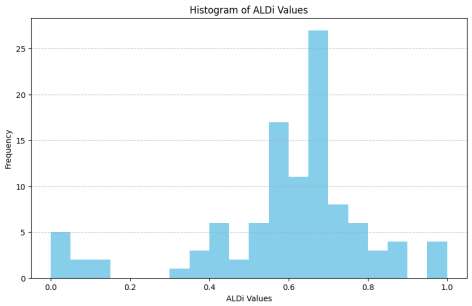

Figure 2: Provided Dev set distribution

4.2 Baseline Architecture

We utilized the Huggingface's library (Wolf et al., 2020) to fine-tune the encoders described in Section 4.1 on a regression task for Sub-task 2 (Abdul-Mageed et al., 2024). Our baseline system, shown in Figure 3 consists of three main components: a transformer-based encoder, customizable regressor layers, and dropout regularization. The BERT encoder serves as the backbone for extracting contextualized representations of input tokens. These representations are then passed through the regressor, which consists of multiple linear layers, dropout layers, and activation functions, to learn non-linear mappings between the input features and the regression target.

This architecture offers flexibility in configurations, such as adjusting the dropout rate and the dimensions of the hidden layers in the regressor.

4.3 Implementation

We utilized the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 5×10^{-6} and an epsilon value of 1×10^{-8} to update model parameters during training. The default RMSE loss function is employed. A dropout rate of 0.2 is applied for regularization. The regressor component consists of consecutive layers: dropout, linear (input size: 768, output size: 512), ReLU (Agarap, 2018) activation, dropout, linear (input size: 512, output size: 256), ReLU activation, and linear (input size: 256, output size: 1). Initially, we applied a sigmoid activation (Dubey et al., 2022) to bound the output between 0 and 1, but this led to a decrease in system performance.

4.4 Weighted Random Sampler (WRS)

The WRS in Torch (Paszke et al., 2019) involves sampling instances from the dataset with probabilities proportional to their weights. In this paper, we used the inverse frequency of the labels to assign these probabilities to each sample. This ensured that instances with lower frequencies, which are more challenging, were given higher probabilities of being selected during training.

4.5 Contrastive Loss

We delve into a different training approach (Pintea et al., 2023) known as contrastive learning (Cole et al., 2022; Shapiro et al., 2022). It aims to reduce the gap in pooled BERT representations between similar pairs while widening the distance between dissimilar pairs. One commonly used contrastive loss is the Normalized Temperature-scaled Cross Entropy (NT-Xent) loss (Chen et al., 2020), which we used in our experiments. Its name is derived from its components. Firstly, the term "Normalized" refers to the utilization of cosine similarity, which generates a standardized score within the range of -1.0 to +1.0. Secondly, "Temperature-scaled" indicates that a temperature parameter adjusts the cosine similarity among all pairs before the computation of the cross-entropy loss. In our training, we try temperature parameters of 0.3 and 0.5 as used by (Chen et al., 2020), and 1. Lastly, the core of this loss function lies in the "Cross-entropy loss" which is fundamentally a multi-class (single-label) cross-entropy loss. In our system, we used the pooled output twice, first passing it through the regressor part of the model, and secondly, passing it to the NT-Xent loss along with the labels during the training process. Integrating the NT-Xent loss with the RMSE loss significantly enhanced our system's performance.

4.6 Ensemble

Ensemble methods in NLP involve combining predictions by leveraging diverse models or variations of the same model to enhance performance. In our system, we apply a brute-force strategy with

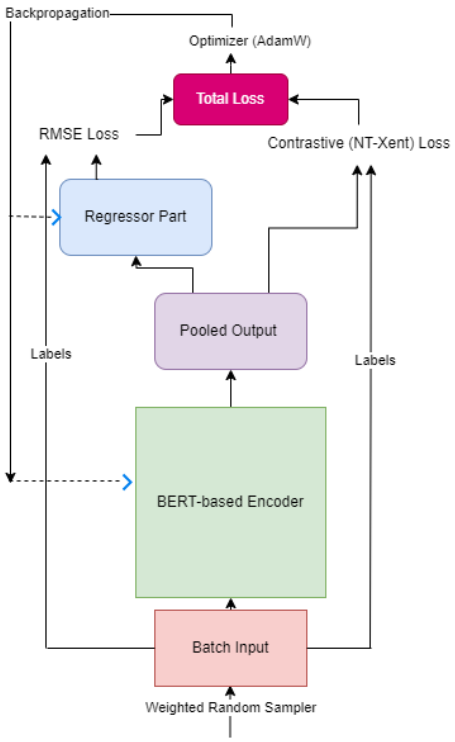

Figure 3: Workflow of Architectural Model Training

many checkpoints coming from different training paradigms, trying to maximize the score on the subtask 2 dev set. This approach includes models with varying temperature parameters in the NT-Xent loss, models trained on the complete training dataset, and models trained after undersampling the majority class (MSA samples) with zero labels. We also experimented with various strategies for aggregating predictions within our ensemble system, such as majority voting, averaging, and median prediction. Among these methods, we found that using the median of the predictions consistently yielded the best results.

5 Results

We saved the best 4 checkpoints on our dev set, which represents 15% from the subtask 2 provided training data, then tested those 4 checkpoints in the shared task dev set to get the best checkpoint. The results on the subtask 2 dev set and test set are given in Table 2. Experiments 1, 2, 3, 5, and 6 determined the use of the AraBERTv2-Twitter-Base encoder in subsequent experiments. Experiments 1 and 4 determined that we would not do any additional preprocessing on the given dataset. Experiment 6 showed improvement when adding WRS to the system. Experiment 7 showed improvement when adding CL to the system compared to Experiment 2. Experiments 8, 9, 10, and 11 showed that combining WRS with CL made further improvements to our system. Experiment 13 showed that our brute-force ensemble strategy when taking the median of the predictions, achieved the best result.

6 Qualitative Results

In this section, we provide practical examples drawn from the dev set of the shared task. Subsequently, we conduct a comparative analysis between the outputs generated by our best system based on the ensemble and the actual ground truth, as detailed in Table 1. Examples numbered 1, and 2 are deemed unsatisfactory instances where our model's predictions were notably inaccurate. In Example 1, the sub-sentence

الله يفضح عرضكم

, may lead to a lower in the dialect score as predicted by our system. In Example 2, the atypical practice of writing English words in an Arabic style, not commonly found in the training set, likely confounds the model. In contrast, Examples 3 and 4 feature words and sub-sentences that frequently appear in the training set, such as

الموضوع، ليه كده، يوم شفت، تذكرت، ماده

, which showcase successful predictions and underscore the effectiveness of our approach.

No	Text	PRD	GT
1	فزرونا الله يفضح عرضكم من بيريز لاصغر لاعب # خزوق # هلا _مدريد	0.576	0.888
2	اقولها لتس فيس تو فيس افتر ديز	0.792	1.0
3	هو الموضوع قالب ع كارديو ليه كده	0.791	0.799
4	يوم شفت الوقت الساعه تذكرت عندي ماده كنت بسحبها ونسيت	0.665	0.666

Table 1: Qualitative Results and Comparison between Prediction (PRD) results and Ground Truth (GT)

Experiment	Model	RMSE	
		Dev Set	Test Set
1	AraBERTv2-Twitter-Large-RMSE	0.1352	-
2	AraBERTv2-Twitter-Base-RMSE	0.1487	-
3	MARBERTv2-RMSE	0.1664	-
4	AraBERTv2-Twitter-Large-RMNALL-RMSE	0.1386	-
5	AraBERTv2-Twitter-Large-WRS-RMSE	0.1366	-
6	AraBERTv2-Twitter-Base-WRS-RMSE	0.1203	-
7	AraBERTv2-Twitter-Base-RMSE-CL (T = 0.5)	0.1368	-
8	AraBERTv2-Twitter-Base-WRS-RMSE-CL (T = 0.3)	0.1138	-
9	AraBERTv2-Twitter-Base-WRS-RMSE-CL (T = 0.5)	0.1113	0.1427
10	AraBERTv2-Twitter-Base-WRS-RMSE-CL (T = 1)	0.1122	-
11	AraBERTv2-Twitter-Base-DS-WRS-RMSE-CL (T = 0.5)	0.1172	-
12	Ensemble Averaging	0.1087	-
13	Ensemble Median	**0.1060**	**0.1406**

Table 2: Results on Subtask 2 Dev Set and Test Set (Leader board Results)

7 Discussion

In Experiment 4, we cleaned the dataset by RMNALL. Surprisingly, this led to a decrease in performance. This decline might be due to inconsistencies among the dataset annotators, who could have considered these elements when scoring the samples. As a result of the reduced performance, we chose not to implement these preprocessing steps.

Considering the imbalanced distribution of the given training set, we sought techniques to mitigate bias towards the majority class. Finally adding the WRS and combining the RMSE with contrastive loss (experiments 6,7, and 9) led to a significant improvement in the results.

Experiments 1, 2, 5, and 6 showed that adding WRS led to an increase in the performance of the AraBERTv2-Twitter base rather than the large version. This is because WRS led to overfitting to the minority samples, so we needed to reduce the complexity of the model.

The noticeable difference between the results of the shared task dev set and test set (leaderboard results) in Experiment 9, 13 in Table 2 may be due to the shared task dev set containing only 107 samples, whereas the blind test set includes 857 samples. Thus, this dev set may not be a good representation of the blind test set.

8 Conclusion

In this paper, we address the challenge of recognizing the nuanced spectrum of dialectness in Arabic text. We developed a sophisticated approach that surpasses binary classification by acknowledging the continuum of dialect variation. By using WRS, CL, and a robust ensemble strategy, our solution achieved a noteworthy RMSE of 0.1406, ranking second on the leaderboard.

References

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. QADI: Arabic dialect identification in the wild. In *Proceed-*

ings of the Sixth Arabic Natural Language Processing Workshop, pages 1–10, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021a. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021b. Arbert marbert: Deep bidirectional transformers for arabic. *Preprint*, arXiv:2101.01785.

Muhammad Abdul-Mageed, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. 2024. NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task. In *Proceedings of The Second Arabic Natural Language Processing Conference (ArabicNLP 2024)*.

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Abien Fred Agarap. 2018. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*.

Maha J. Althobaiti. 2020. Automatic arabic dialect identification systems for written texts: A survey. *CoRR*, abs/2009.12622.

Maha J. Althobaiti. 2022. Creation of annotated country-level dialectal arabic resources: An unsupervised approach. *Natural Language Engineering*, 28(5):607–648.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Elijah Cole, Xuan Yang, Kimberly Wilber, Oisin Mac Aodha, and Serge Belongie. 2022. When does contrastive visual representation learning work? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14755–14764.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shiv Ram Dubey, Satish Kumar Singh, and Bidyut Baran Chaudhuri. 2022. Activation functions in deep learning: A comprehensive survey and benchmark. *Preprint*, arXiv:2109.14545.

Heba Elfardy and Mona Diab. 2013. Sentence level dialect identification in Arabic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 456–461, Sofia, Bulgaria. Association for Computational Linguistics.

Nizar Y Habash. 2010. *Introduction to Arabic natural language processing*. Morgan & Claypool Publishers.

Amr Keleg, Sharon Goldwater, and Walid Magdy. 2023. ALDi: Quantifying the Arabic level of dialectness of text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10597–10611, Singapore. Association for Computational Linguistics.

Amr Keleg and Walid Magdy. 2023. Arabic dialect identification under scrutiny: Limitations of single-label classification. In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *Preprint*, arXiv:1711.05101.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49, Austin, Texas. Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

Silvia L Pintea, Yancong Lin, Jouke Dijkstra, and Jan C van Gemert. 2023. A step towards understanding

why classification helps regression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19972–19981.

A. Sakr and M. Torki. 2023. Arapunc: Arabic punctuation restoration using transformers. In *2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, pages 1–6, Los Alamitos, CA, USA. IEEE Computer Society.

Mohammad Salameh, Houda Bouamor, and Nizar Habash. 2018. Fine-grained Arabic dialect identification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1332–1344, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ahmad Shapiro, Ayman Khalafallah, and Marwan Torki. 2022. AlexU-AIC at Arabic hate speech 2022: Contrast to classify. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 200–208, Marseille, France. European Language Resources Association.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Omar F. Zaidan and Chris Callison-Burch. 2011. The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 37–41, Portland, Oregon, USA. Association for Computational Linguistics.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.